International Journal of Innovative Research in Computerand Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

||Volume 8, Issue 5, May 2020||

Predicting the Accuracy of Diabetes Using Different Mining Technique

Harini S [1], Dr A Gnanabaskaran[2]

B.E Student, Dept. of CSE, K.S. Rangasamy College of Technology, Tiruchengode, Tamilnadu, India[1]

Professor, K.S. Rangasamy College of Technology, Tiruchengode, Tamilnadu, India[2]

ABSTRACT: Due to its continuously increasing occurrence, more and more families are influenced by diabetes mellitus. Most diabetics know little about their health quality or the risk factors they face prior to diagnosis. In this study, we have proposed a novel model based on data mining techniques for predicting type 2 diabetes mellitus (T2DM). The main problems that we are trying to solve are to improve the accuracy of the prediction model, and to make the model adaptive to more than one dataset. Based on a series of pre-processing procedures, the model is comprised of two parts, the improved K-means algorithm and the logistic regression algorithm.

The Pima Indians Diabetes Dataset and the Waikato Environment for Knowledge Analysis toolkit were utilized to compare our results with the results from other researchers. The conclusion shows that the model attained a 3.04% higher accuracy of prediction than those of other researchers. Moreover, our model ensures that the dataset quality is sufficient. To further evaluate the performance of our model, we applied it to two other diabetes datasets. Both experiments' results show good performance. As a result, the model is shown to be useful for the realistic health management of diabetes.

I. INTRODUCTION

1.1 PREDICTIVE MODELING

Regression as technique although is predictive technique, but based on analyzes conducted to reach the conclusion most scientists, they have concluded that the reliability percentage is around 95%. Predictive modeling is a name given to a collection of mathematical techniques having in common the goal of finding a mathematical relationship between a target, response, or “dependent” variable and various predictor or “independent” variables with the goal in mind of measuring future values of those predictors and inserting them into the mathematical relationship to predict future values of the target variable.

Regression analysis establishes a relationship between a dependent or outcome variable and a set of predictors. Regression, as a data mining technique, is supervised learning. Supervised learning partitions the database into training and validation data. The techniques used in this research were simple linear regression and multiple linear regression. Some distinctions between the use of regression in statistics verses data mining are: in statistics The data is a sample from a population , but in Data Mining The data is taken from a large database (e.g. 1 million records). Also in statistics.

The preceding view shows data mining as one step in the knowledge discovery process, albeit an essential one because it uncovers hidden patterns for evaluation. However, in industry, in media, and in the research milieu, the term data mining is often used to refer to the entire knowledge discovery process (perhaps because the term is shorter than knowledge discovery from data).

The development of Information Technology has generated large amount of databases and huge data in various areas. The research in databases and information technology has given rise to an approach to store and manipulate this precious data for further decision making. Data mining is a process of extraction of useful information and patterns from huge data. It is also called as knowledge discovery process, knowledge mining from data, knowledge extraction or data/pattern analysis.

Regression technique can be adapted for predication. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. In data mining independent variables are attributes already known and response variables are what we want to predict. Unfortunately, many real-world problems are not simply prediction. For instance, sales volumes, stock prices, and product failure rates are all very difficult to predict because they may depend on complex interactions of multiple predictor variables.

Types of regression methods:

- Linear Regression
- Multivariate Linear Regression
- Nonlinear Regression

- Multivariate Nonlinear Regression

Neural network is a set of connected input/output units and each connection has a weight present with it. During the learning phase, network learns by adjusting weights so as to be able to predict the correct class labels of the input tuples. Neural networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. These are well suited for continuous valued inputs and outputs. For example handwritten character reorganization, for training a computer to pronounce English text and many real world business problems and have already been successfully applied in many industries. Neural networks are best at identifying patterns or trends in data and well suited for prediction or forecasting needs. Types of neural networks: Back Propagation.

The simple linear regression model then could be viewed as the line that minimized the error rate between the actual prediction value and the point on the line (the prediction from the model). The simplest form of regression seeks to build a predictive model that is a line that maps between each predictor value to a prediction value Fraud detection and credit risk applications are particularly well suited to this type of analysis. This approach frequently employs decision tree or neural network- based classification algorithms. The data classification process involves learning and classification. In Learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuple.

1.2 DATA WAREHOUSING

Data warehousing is a collection of methods, techniques, and tools used to support knowledge workers—senior managers, directors, managers, and analysts—to conduct data analyses that help with performing decision-making processes and improving information resources.

Data warehousing is a phenomenon that grew from the huge amount of electronic data stored in recent years and from the urgent need to use that data to accomplish goals that go beyond the routine tasks linked to daily processing. In a typical scenario, a large corporation has many branches, and senior managers need to quantify and evaluate how each branch contributes to the global business performance. The corporate database stores detailed data on the tasks performed by branches. To meet the managers' needs, tailor-made queries can be issued to retrieve the required data. In order for this process to work, database administrators must first formulate the desired query (typically an aggregate SQL query) after closely studying database Then the query is processed. This can take a few hours because of the huge amount of data, the query complexity, and the concurrent effects of other regular workload queries on data. Finally, a report is generated and passed to senior managers in the form of a spreadsheet.

1.2.1 Decision Support System

A decision support system (DSS) is a set of expandable, interactive IT techniques and tools designed for processing and data and for supporting managers in decision making. To do this, the system matches individual resources of managers with computer resources to improve the quality of the decisions made.

An exponential increase in operational data has made computers the only tools suitable for providing data for decision-making performed by business managers. This fact has dramatically affected the role of enterprise databases and fostered the introduction of decision support systems. The concept of decision support systems mainly evolved from two research fields: theoretical studies on decision-making processes for organizations and technical research on interactive IT systems. However, the decision support system concept is based on several disciplines, such as databases, artificial intelligence, man-machine interaction, and simulation.

Data warehousing is a collection of methods, techniques, and tools used to support knowledge workers—senior managers, directors, managers, and analysts—to conduct data analyses that help with performing decision-making processes and improving information resources. A data warehouse should provide a unified view of all the data. Generally speaking, we can state that creating a data warehouse system does not require that new information be added; rather, existing information needs rearranging. This implicitly means that an information system should be previously available.

Operational data usually covers a short period of time, because most transactions involve the latest data. A data warehouse should enable analyses that instead cover a few years. For this reason, data warehouses are regularly updated from operational data and keep on growing. If data were visually represented, it might progress like so: A photograph of operational data would be made at regular intervals

1.2.2 Data Warehouse Architectures.

- **Separation:** Analytical and transactional processing should be kept apart as much as possible.
- **Scalability:** Hardware and software architectures should be easy to upgrade as the data volume, which has to be managed and processed, and the number of users' requirements, which have to be met, progressively increase.

- **Extensibility:** The architecture should be able to host new applications and technologies without redesigning the whole system.
- **Security:** Monitoring accesses is essential because of the strategic data stored in data warehouses.
- **Administerability** Data warehouse management should not be overly difficult.

1.2.4 Single-Layer Architecture

A single-layer architecture is not frequently used in practice. Its goal is to minimize the amount of data stored; to reach this goal, it removes data redundancies. Figure 1-2 shows the only layer physically available: the source layer. In this case, data warehouses are *virtual*.

A data mart is a subset or an aggregation of the data stored to a primary data warehouse. It includes a set of information pieces relevant to a specific business area, corporate department, or category of users.

OLAP might be the main way to exploit information in a data warehouse. Surely it is the most popular one, and it gives end users, whose analysis needs are not easy to define beforehand, the opportunity to analyze and explore data interactively on the basis of the multidimensional model. While users of reporting tools essentially play a passive role, OLAP users are able to start a complex analysis session actively, where each step is the result of the outcome of preceding steps. Real-time properties of OLAP sessions, required in-depth knowledge of data, complex queries that can be issued, and design for users not familiar with IT make the tools in use play a crucial role. The GUI of these tools must be flexible, easy-to-use, and effective.

In many applications, an intermediate approach between static reporting and OLAP is broadly used. This intermediate approach is called semi-static reporting. Even if a semi-static report focuses on a group of information previously set, it gives users some margin of freedom. Thanks to this margin, users can follow a limited set of navigation paths.

Transformation is the core of the reconciliation phase. It converts data from its operational source format into a specific data warehouse format. If you implement a three-layer architecture, this phase outputs your reconciled data layer. Independently of the presence of a reconciled data layer, establishing a mapping between the source data layer and the data warehouse layer is generally made difficult by the presence of many different, heterogeneous sources. If this is the case, a complex integration phase is required when designing your data warehouse.

1.3 mHEALTH

Health care organizations – providers, payers, and life sciences companies – should consider each of the four dimensions of mHealth as they weigh market entry.

- **Personalize the consumer's experience.** mHealth offers tools – consumer engagement strategies, retail capabilities based upon mobile platforms and data analytics, digitization of an individual's wellness and health care needs, and more – that can enable competitive differentiation by creating personalized solutions that help drive consumer loyalty.
- **Keep it simple.** mHealth functionality should be easy to use and akin to users' mobile experiences in hospitality, retail, travel, and banking.
- **Pay attention to privacy and security.** mHealth technologies and permeable boundaries among existing and new entrants in the health care ecosystem increase the complexity of managing protected health information (PHI). Organizations need to be secure, vigilant, and resilient in the face of threats to information security.

mHealth's communication, coordination, monitoring, and data collection capabilities may provide tools to help achieve the goals inherent in new service delivery models such as accountable care organizations. These new delivery models an reimbursement reforms such as bundled payments are predicated on achieving quality outcomes against evidence-based standards. Payments are based on evaluations of clinical processes, patient experiences, outcomes, and efficiencies against industry benchmarks and base thresholds, including incorporating data flowing both remotely and at the point of care.

mHealth technologies and permeable boundaries among existing and new entrants in the health ecosystem increase the complexity of managing PHI and compound an already challenging issue for industry stakeholders.

II. EXISTING SYSTEM

It has classified only undesirable effect of changing a Diabetes patient's existing test data groups, potentially undoing the patient's own manual efforts in organizing her history. It involves a high computational cost, have to repeat a large number of attribute test data group similarity computations for every new test data.

As existing approaches to extract Diabetes Disease prediction suffer from scalability. It is imperative to address the scalability issue. Connections in Diabetes prediction are not homogeneous.

Diabetes infection has endangered 2.5 billion populations all around the world. Every year there are 50 million people who suffer from it globally [1]. Pakistan has been victim of this rapidly growing sickness from last few years. Since 2007 in Pakistan, large number of cases was marked especially in Lahore. In 1994 at Karachi Pakistan's first case of Diabetes was appeared and Diabetes's outbreak in 2011, that was more life-threatening than preceding years and 1400 people were affected.

2.1 DRAWBACKS

Motivate and propose a method to perform test data grouping in a dynamic fashion. Our goal is to ensure good performance while avoiding disruption of existing patient-defined test data groups

- Improper classification may provide wrong results
- Poor performance
- Complex data processing to find Diabetes prediction
- The data retrieval based on user requirement is not done
- This relation-type information, however, is often not readily available in Diabetes prediction

III. PROPOSED SYSTEM

Methods that can accurately predict Diabetes Disease are greatly needed and good prediction techniques can help to predict Diabetes Disease more accurately. In this system, it used two feature selection methods, forward selection (FS) and backward selection (BS), to remove irrelevant features for improving the results of Diabetes Disease prediction. The results show that feature reduction is useful for improving the predictive accuracy and density is irrelevant feature in the dataset where the data had been identified on full field digital diabetes collected at the UCI Repository. In addition, decision tree (DT), support vector machine sequential minimal optimization (SVM-SMO) and their ensembles were applied to solve the Diabetes Disease diagnostic problem in an attempt to predict results with better performance. The results demonstrate that ensemble classifiers are more accurate than a single classifier.

The proposed framework SMO based on disease prediction is shown to be effective in addressing this prediction. The framework suggests a novel way of network classification: first, capture the latent affiliations of actors by extracting disease prediction based on network connectivity, and next, apply extant data mining techniques to classification based on the extracted prediction.

In the initial study, modularity maximization was employed to extract disease prediction. The superiority of this framework over other representative relational learning methods has been verified with Diabetes prediction Diabetes data.

Prove that with this proposed approach, sparsity of disease prediction is guaranteed.

3.1 SVM

We aim to find a feature subset of size m which contains the most informative features. The two well performing feature selection algorithms on the dataset are briefly outlined below.

Feature reduction applies a mapping of the multidimensional space into a space of lower dimensions. Feature extraction includes features construction, space dimensionality reduction, sparse representations, and feature selection all these techniques are commonly used as preprocessing to machine learning and statistics tasks of prediction, including pattern recognition. Although such problems have been tackled by researchers for many years, there has been recently a renewed interest in feature extraction. The feature space having reduced features truly contributes to classification that cuts preprocessing costs and minimizes the effects of the „peaking phenomenon" in classification. Thereby improving the overall performance of classifier based intrusion detection systems. SVM is a linear transformation with linear ortho normal basis vectors; it can be expressed by a translation and rotation. The below figures mention the snapshots of implementation process where first image represents the preprocessing stage and third diagram denotes the decision tree generated with Fisher filtering

3.2 ADVANTAGES

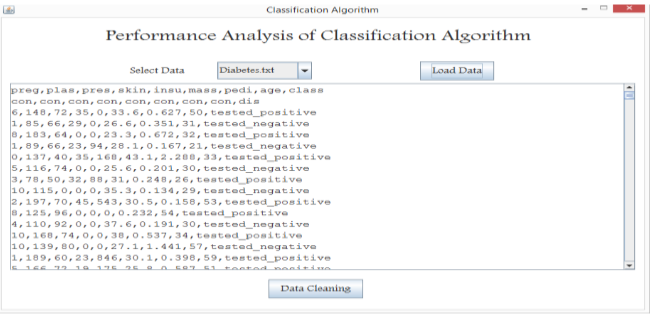
High effectiveness of the proposed algorithms in capturing test data relevance.

- Test data reformulation graph and the test data click graph into a single graph that it refer to as the test data fusion graph, and by expanding the test data set when classification relevance occur.
- High Relevance Measure
- Good classification accuracy
- Online test data grouping process. High Similarity function provides types of Diabetes data classification
- It supports multiple disease classification with better accuracy.

IV. MODULES

4.1 DATA VISUALIZATION AND PRE-PROCESSING

The Wisconsin Prognostic Diabetes Disease dataset is downloaded from the UCI Machine Learning Repository website and saved as a text file. This file is then imported into Excel spreadsheet and the values are saved with the corresponding attributes as column headers. The missing values are replaced with appropriate values. The ID of the patient cases does not contribute to the classifier performance. Hence it is removed and the outcome attribute defines the target or dependent variable thus reducing the feature set size to 33 attributes. The algorithmic techniques applied for feature relevance analysis and classification are elaborately presented in the following sections.



```
preg,plas,pres,skin,insu,mass,pedi,age,class
con,con,con,con,con,con,con,con,dis
6,148,72,35,0,33.6,0.627,50,tested_positive
1,85,66,29,0,26.6,0.351,31,tested_negative
8,183,64,0,0,23.3,0.672,32,tested_positive
1,89,66,23,94,28.1,0.167,21,tested_negative
0,137,40,35,168,43.1,2.288,33,tested_positive
5,116,74,0,0,25.6,0.201,30,tested_negative
3,78,50,32,88,31,0.248,26,tested_positive
10,115,0,0,0,35.3,0.134,29,tested_negative
2,197,70,45,543,30.5,0.158,53,tested_positive
8,125,96,0,0,0,0.232,54,tested_positive
4,110,92,0,0,37.6,0.191,30,tested_negative
10,168,74,0,0,38,0.537,34,tested_positive
10,139,80,0,0,27.1,1.441,57,tested_negative
1,189,60,23,846,30.1,0.398,59,tested_positive
```

4.1.1 PERFORMNCE

4.2 SMO ATTRIBUTE FEATURE SELECTION ALGORITHMS

The generic problem of supervised feature selection can be outlined as follows. Given a data set {(xi, yi)} ni = 1 where xi ∈ Rd and yi ∈ {1, 2…c}, we aim to find a feature subset of size m which contains the most informative features. The two well-performing feature selection algorithms on the WPDC dataset are briefly outlined below.

4.2.1 Mean and STD Score Filtering

It is termed Univariate Mean and STD Score's ANOVA ranking. It is a supervised feature selection algorithm that processes the selection independently from the learning algorithm. It follows a filtering approach that ranks the input attributes according to their relevance. A cutting rule enables the selection of a subset of these attributes. It is required to define the target attribute which in this domain of research applies to the nature of the Diabetes Disease (recurrent/non- recurrent) and the predictor attributes. After computing the Mean and STD Score for each feature, it selects the top-m ranked features with large scores..

4.3 LEVERAGE BACKWARD LOGISTIC REGRESSION RISK ANALYSIS

When the number of descriptors is very large for a given problem domain, a learning algorithm is faced with the problem of selecting a relevant subset of features backward regression includes regression models in which the choice of predictor variables is carried out by an automatic procedure. The iterations of the algorithm for logistic regression are given in steps as stated as follows.

Step 1: The feature set with all 'ALL' predictors.

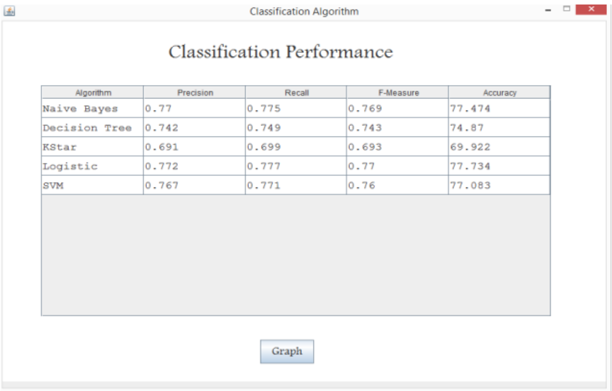
Step 2: Eliminate predictors one by one.

Step 3: 'ALL' models are learnt containing 'ALL-1' descriptor each.

These iterations are further continued till either a pre-specified target size is reached or the desired performance statistics (classification accuracy) is obtained. After feature relevance, it classify the nature of the Diabetes Disease cases in the Wisconsin Prognostic Diabetes Disease dataset using twenty classification algorithms. The best performing algorithms are described in the following section.

4.4 FEATURE REDUCTION BY SMO

Feature reduction applies a mapping of the multidimensional space into a space of lower dimensions. Feature extraction includes features construction, space dimensionality reduction, sparse representations, and feature selection all these techniques are commonly used as preprocessing to machine learning and statistics tasks of prediction, including pattern recognition. Although such problems have been tackled by researchers for many years, there has been recently a renewed interest in feature extraction. The feature space having reduced features truly Contributes to classification that cuts preprocessing costs and minimizes the effects of the 'peaking phenomenon' in classification. Thereby improving the overall performance of classifier based intrusion detection systems. The commonly used dimensionality reduction methods include supervised approaches such as Linear Discriminant Analysis (LDA), unsupervised ones such as SMO, and additional spectral and manifold learning methods. It converts a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components is less than or equal to the number of original variables.



Classification Performance

Algorithm	Precision	Recall	F-Measure	Accuracy
Naive Bayes	0.77	0.775	0.769	77.474
Decision Tree	0.742	0.749	0.743	74.87
KStar	0.691	0.699	0.693	69.922
Logistic	0.772	0.777	0.77	77.734
SVM	0.767	0.771	0.76	77.083

4.3.3CLASSIFICATION PERFORMANCE

V. CONCLUSION

The research test different algorithms. The result of the research focused on correctness of the algorithms in the training. It depended on WDBC data set. The test result shows that the SMO is the best algorithm. The best way was when the research removed the sample for missing value in training for SMO. However, Random Tree result was keep better correctness when keeps the sample for missing value.

The research undertook an experiment on application of various data mining algorithms to predict the Diabetes and to compare the best method of prediction. The research results do not present dramatic differences in the prediction when using different classification algorithms in data mining.

The experiment can serve as an important tool for physicians to predict risky cases in the practice and advise accordingly.

The model from the classification will be able to answer more complex queries in the prediction of Diabetes diseases.

The predictive accuracy determined by SMO algorithm suggests that parameters used are reliable indicators to predict the presence of Diabetes diseases.

REFERENCES

[1] Aruna Pavate and Nazneen Ansari,(2017)," Risk Prediction of Disease Complications in Type 2 Diabetes Patients Using Soft Computing Techniques", Fifth International Conference on Advances in Computing and Communications.

[2] Gang Shi, Shanshan Liu and Ding Ye, (2018)," Design and Implementation of Diabetes Risk Assessment Model Based On Mobile Things", 7th International Conference on Information Technology in Medicine and Education.

[3] Han Longfei, Luo Senlin, (2017)" An intelligible risk stratification model based on pairwise and size constrained Kmeans"., IEEE J Biomed Health Inf 2017;21(5):1288–96.

[4] Juntao Wang and Xiaolong Su, (2019)" An improved K-Means clustering algorithm", IEEE 3rd International Conference on Communication Software and Networks (ICCSN).

[5] Li Huan, Zhang Qi, Lu Kejie, (2016)" Integrating mobile sensing and social network for personalized health-care application" Health care information systems; 2016.

[6] Marcano-Cede~no Alexis, Torres Joaquín, Andina Diego (2017)" A prediction model to diabetes using artificial metaplasticity", IWINAC 2011, Part II. LNCS 6687; 2011. p. 418–25.

[7] Md Abul Basar, Hassan Nomani Alvi, Gazi, (2018)" A Review on Diabetes Patient Lifestyle Management Using Mobile Application", 18th International Conference on Computer and Information Technology (ICCIT), 21-23 December, 2018.

[8] Mechelle Gittens, Reco King, Curtis Gittens and Adrian Als (2016)" Post-diagnosis Management of Diabetes through a Mobile Health Consultation Application", 2016 IEEE 16th International Conference on e-Health Networking, Applications and Services (Healthcom).

[9] Michie D, Spiegelhalter DJ, Taylor CC, (2017)" Machine learning, neural and statistical classification". Ellis Horwood.

[10] Muhammad H. Aboelfotoh, Patrick Martin and Hossam S. Hassanein, (2017)"A mobilebased architecture for integrating personal health record data",IEEE 16th International Conference on e-Health Networking, Applications and Services (Healthcom).

[11] Naganna Chetty P, (2016)" An Improved Method for Disease Prediction using Fuzzy Approach",Second International Conference on Advances in Computing and Communication Engineering.

[12] Phattharat Songthung and Kunwadee Sripanidkulchai, (2018)" Improving Type 2 Diabetes Mellitus Risk Prediction Using Classification", 2018 13th International Joint Conference on Computer Science and Software Engineering (JCSSE).

[13] Qasim Majeed, Hayder Hbail and Abdolah Chalechale, (2017)" A Comprehensive Mobile EHealthcare System", IKT2015 7th International Conference on Information and Knowledge Technology.

[14] Blaz Z, (2017)" Predictive data mining in clinical medicine: current issues and guidelines". Int J Med Inf 2008;77:81–97.

[15] Rojalina Priyadarshini, Nilamadhab Dash and Rachita Mishra, (2017)" A Novel approach to Predict Diabetes Mellitus using Modified Extreme Learning Machine.".

[16] Schnall Rebecca, Rojas Marlen, (2017) "A user-centered model for designing consumer mobile health (mHealth) applications (apps)", J Biomed Inf 2016;60:243–51.

[17] Shunye Wang, (2017)" Improved K-means clustering algorithm based on the optimized initial centroids", 2017 3rd International Conference on Computer Science and Network Technology (ICCSNT).

[18] Sowjanya K, Mob DB, (2017)" A machine learning based system for predicting diabetes risk using mobile devices", IEEE International Advance Computing Conference (IACC).